CS 7545: Machine Learning Theory **Fall 2021**

Lecture 6

Name: Tyler LaBonte **Professor:** Santosh S. Vempala

1 Robust Estimation

Suppose we obtain a $1-\epsilon$ fraction of our sample from the true model and an ϵ fraction from an adversary. Can we still learn or estimate the true model?

Clearly the sample mean is nonrobust since the adversary can perturb points arbitrarily far. Similarly, for singular vectors $\boldsymbol{v}$, we can move a point orthogonal to $\boldsymbol{v}$ and really far away to completely change $\boldsymbol{v}$. Thus, low-degree sample moments are not robust estimators.

Consider a k-GMM with ϵ-adversarial corruption. All the methods we used so far depend on low-degree moments, so we need a new strategy. Consider a simpler version to start with – the one-dimensional Gaussian $\mathcal{N}(\mu, \sigma^2)$.

Claim 1.1

The sample median has

$$|\mu - \text{med}| \leq \epsilon\sigma, \tag{1}$$

since if you move ϵ fraction of points to the other side of the mean, the median will move at most $\epsilon\sigma$. In fact, the median is the best possible robust estimator information-theoretically, in the sense that it allows us to tell apart $\mathcal{N}(0,1)$ and $\mathcal{N}(\epsilon,1)$.

In higher dimension, the coordinate-wise median does not work since the error scales with the dimension. One idea is to define a minimum-volume ellipsoid containing half the points and output its center; however, computing this ellipsoid is **NP**-hard in general.

In 2016, a method was found for estimating mean and covariance to within dimension-free, information-theoretic bounds. For $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d)$ we have $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 = \mathcal{O}(\epsilon\sqrt{\log 1/\epsilon})$, and with only additive corruptions (*i.e.,* the data poisoning model) the bound improves to $O(\epsilon)$. The bound is the same for general Gaussians except we replace $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2$ with $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|_2$.

Theorem 1.1: Robust Mean Estimation

Suppose $G = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. If the sample covariance $\hat{\boldsymbol{\Sigma}}$ satisfies $\|\hat{\boldsymbol{\Sigma}}\|_2 \leq 1 - \epsilon$, then $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = \mathcal{O}(\epsilon\sqrt{\log 1/\epsilon})$ (and $\mathcal{O}(\epsilon)$ for additive corruption). The intepretation is that $\|\hat{\boldsymbol{\Sigma}}\|_2 \leq 1 - \epsilon$ implies that the largest eigenvalue, which is the largest variance in any direction, is ϵ-close to the true variance. Since this is true for the largest eigenvalue, it is true for all eigenvalues; so, if we can estimate the variance in all directions ϵ-close, we obtain the bound.

Proof: Suppose we are in the additive corruption scenario, so $S = G \cup B$. Then $\hat{\boldsymbol{\mu}} = \epsilon \boldsymbol{\mu}_B$. Recall

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\intercal - \boldsymbol{\mu}\boldsymbol{\mu}^\intercal; \tag{2}$$

expanding this for additive corruption we obtain

$$\hat{\boldsymbol{\Sigma}} = (1-\epsilon)\mathbf{I} + \epsilon \boldsymbol{\Sigma}_B + (\epsilon - \epsilon^2)\boldsymbol{\mu}_B \boldsymbol{\mu}_B^\intercal. \tag{3}$$

Since $1 + \epsilon \geq \boldsymbol{v}^\intercal \hat{\boldsymbol{\Sigma}} \boldsymbol{v}$ by assumption and $\boldsymbol{v} = \frac{\boldsymbol{\mu}_B}{\|\boldsymbol{\mu}_B\|}$ since the greatest change of variance is in the direction of the noise, we have

$$\hat{\boldsymbol{\Sigma}} = (1-\epsilon) + (\epsilon - \epsilon^2)\|\boldsymbol{\mu}_B\|^2 \tag{4}$$

Thus

$$1 \geq \frac{1-\epsilon}{\epsilon^2} \|\hat{\boldsymbol{\mu}}\|, \tag{5}$$

so $\|\hat{\boldsymbol{\mu}}\| = \mathcal{O}(\epsilon)$. ■

Lemma 1.1

Suppose we remove $\boldsymbol{x}_i$ with $\|\boldsymbol{x}_i\| > c\sqrt{d}$ as they are likely noise. Then from Theorem 1.1 we have that $\lambda_{\min}(\hat{\boldsymbol{\Sigma}}) \geq 1 - \epsilon$ and $\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}) \leq (1 + \mathcal{O}(\epsilon))d$. This implies that $\lambda_{d/2}(\hat{\boldsymbol{\Sigma}}) \leq 1 + \mathcal{O}(\epsilon)$. In other words, the sample mean is a good estimator in the $d/2$ subspace.

This lemma gives an algorithm for robust mean estimation:

1. Remove $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i\| > c\sqrt{d}$.
2. Find an eigendecomposition of $\hat{\boldsymbol{\Sigma}}$. Suppose V is the top $d/2$ subspace and W is the bottom $d/2$ subspace. In W, by the lemma, $\max \|\hat{\boldsymbol{\mu}}_W - \boldsymbol{\mu}_W\| = \mathcal{O}(\epsilon)$.
3. Recurse on V.

This algorithm runs for $\log(d)$ iterations and has error $\mathcal{O}(\epsilon\sqrt{\log d})$, but it is conjectured to have $\mathcal{O}(\epsilon)$ error. There is also another method in the course notes which iteratively deletes large points along high-variance directions.